CA Journal of Cultural Analytics

Topic Modeling, Epistemology, and the English and German Novel

Matt Erlin

05.01.17

According to Rita Felski, context is overrated.[12] Even in the sophisticated variants of contextualization typical of the New Historicism, she explains, scholars' obsession with historical context as the ultimate source of textual meaning disregards the capacity of literature to resonate across time and space. "Why is it," she writes, "that we can feel solicited, button-holed, stirred up, by words that were drafted eons ago?" (576). Felski is not the first to raise such objections. In an essay from 2001, Russell Berman takes a similar approach to the politics of periodization, pointing out how periodizing a work can serve to discipline it, that is to say, to deny its claim on the reader's present. For Berman, "A literary-critical culture that values historical frames over 'artistic pleasure,' . . . tends to dismiss the diachronic moment in any reading, and with it the potential of tradition, the ca-

1

pacity precisely to transcend the constraints of the isolated historical moment."[1] Both Felski and Berman present powerful arguments against the fetishization of contemporaneity as the source of the truth of a work. In their focus on affect, however, they neglect a crucial facet of the transtemporal resonance they seek to underscore. To limit oneself to the question of how and why literature can "solicit," "button-hole," and "stir us up" is to understate its role as both an instrument of cognition and a means of fostering particular cognitive capacities, or at least to sidestep the constitutive role of cognition in the experience of aesthetic pleasure.

The following essay is conceived as a computationally-assisted contribution to thinking about transtemporal—and transnational—categories for understanding literature, and the novel in particular. I take my cue from such scholars as Felski and Berman, but rather than emphasizing affective connections, my focus is on literature as an imaginative space for raising and working through questions about epistemology: the sources, scope, and legitimacy of human knowledge. Interest in the novel and epistemology has already given rise to distinguished scholarship on the part of scholars of both Anglo-American and German literature, the two national traditions under consideration here. Michael McKeon, to name one of the more notable Anglo-American critics, has traced the birth of the modern novel in the early modern period back to an "epistemological crisis" triggered by the materialism and empiricism of an increasingly secular age.[2] More recently, John Bender has made a strong claim for the eighteenth-century novel as absolutely central to the epistemological paradigm shift that occurs in the period.[3] The topic has arguably garnered even more attention from German scholars, though they tend to be less preoccupied with the "rise" of the genre than their Anglo-American counterparts. One ambitious recent example is Ernst-Wilhelm Händler's *Versuch über den Roman als Erkenntnisinstrument* (2014; Essay on the Novel as an Instrument of Cognition), in which the author begins his analysis with the simple assertion "literature generates knowledge that science cannot produce."[4]

---

[1] Russell A. Berman, "Politics: Divide and Rule," *MLQ: Modern Language Quarterly* 62.4 (2001): 328.

[2] Michael McKeon, *The Origins of the English Novel 1600-1740* (Baltimore: Johns Hopkins University Press, 1987), 65-89.

[3] John Bender, "Novel Knowledge: Judgement, Experience, Experiment," in *This is Enlightenment,* ed. Clifford Siskin and William Warner (Chicago: University of Chicago Press, 2010), 290. According to Bender, not only are questions "about method and the nature of knowledge . . . intrinsic to modernity as it takes form during the seventeenth and eighteenth centuries"; he also asserts that the "new novel of the eighteenth century" constitutes "one of the strands in these debates" as well as "one of the modes of experimentation."

[4] Ernst-Wilhelm Händler, *Versuch über den Roman als Erkenntnisinstrument* (Frankfurt/Main: S. Fischer, 2014), 5.

As Händler's statement suggests, epistemological inquiries can be (and have been) addressed to any number of literary genres, not just to the novel. Even in those studies that take a broader approach to the topic, however, the novel often has pride of place, and for good reason.[5] As the most capacious and unruly genre, frequently associated with depictions of the totality of human experience, the novel would seem to instantiate a set of epistemological assumptions in its very structure. Indeed, many studies engage at precisely this structural level, uncovering subterranean epistemological relevance on the basis of finely-grained readings of individual works. In contrast, the analysis that follows begins with a wager on the productivity of a "flat" reading, one that, at least initially, directs our attention to surfaces rather than depths and corpora rather than the individual case study.[6] I proceed from the assumption that we can create a meaningful proxy for epistemological engagement by identifying a set of relevant linguistic features and determining which novels possess these features to the highest degree. While this approach no doubt entails a simplification vis-à-vis some previous readings of novelistic epistemologies, the results to which it gives rise suggest that more subtly hermeneutic approaches may themselves underestimate just how significant explicit and detailed reflections on knowledge are to the cultural work done by the genre. As will become clear, moreover, these same results also point us toward indicators of less conspicuous modes of epistemological inquiry.

In sum, the analysis that follows is conceived as an alternative to the depth-hermeneutic methods that Stephen Best and Sharon Marcus have grouped together under the rubric of "symptomatic reading," but one that also resists the reduction of computational analysis to matters of empirical verification.[7] It is inspired in part by Steven Ramsay's conception of an "algorithmic criticism," understood as a form of criticism that uses computation to foster the "unfolding of interpretive possibilities" rather than as an ostensibly objective means to refute or confirm claims based on more traditional interpretive practices.[8] Computational and statistical methods are not employed here in order to test a particular hypothesis about epistemology and the novel, though such an

---

[5] One example is *Literatur und Wissen: Ein interdisziplinäres Handbuch* (Literature and Knowledge: An Interdisciplinary Handbook), The handbook includes a section of 23 "exemplary readings," 14 of which address novels. Roland Borgards et al (eds), *Literatur und Wissen: Ein interdisziplinäres Handbuch* (Stuttgart: Metzler, 2013).

[6] Heather Love, "Close but not Deep: Literary Ethics and the Descriptive Turn," *New Literary History* 41, no. 2 (2010): 375.

[7] See Stephen Best and Sharon Marcus, "Surface Reading: an Introduction," *Representations* 108, no. 1 (2009): 3.

[8] Steven Ramsey, *Reading Machines: Toward an Algorithmic Criticism.* (Urbana-Champaign, IL: University of Illinois Press, 2011), 10-11.

approach could certainly lead to valuable insights. The aim, rather, is to consider whether such methods can generate new ways of thinking about how novels engage with epistemological questions and how that engagement develops over the course of the long nineteenth century. With regard to the latter question, the preliminary results raise the intriguing possibility of a reorientation in the late nineteenth century, an evolution in which overt philosophical themes are increasingly combined with more subtle epistemological narrative positionings.

Corpus, Approaches, and Initial Results

The text base for the analysis consisted of four corpora in two languages. Our research team first assembled two literary corpora, one a collection of 261 German novels published between 1731 and 1932 and the other containing 262 English and American novels published between 1719 and 1930. These were based primarily on two pre-existing collections originally digitized by the German company Directmedia Publishing, which were then supplemented with additional works from Project Gutenberg to even out the chronological coverage. The novels included are a mix of works most would consider today to be canonical and works that are no longer read frequently but were popular around the time they were published. To each of these we added a smaller philosophy corpus. The philosophy corpora comprised identical selections of excerpts, taken from a well-respected anthology, from twelve canonical works of epistemology in either the original language or in translation. In other words, the German philosophy corpus included either original German texts or previously published German translations, and the English philosophy corpus included those same texts either in their original English or in English translations. Of the twelve excerpts, five appeared originally in English, five in German, and two in French. Their original publication dates range from 1641 to 1950.[9] To establish which novels shared linguistic features with this philosophy corpus, we adopted two distinct techniques.

---

[9]The selection was based on David Cooper, ed., *Epistemology: The Classic Readings* (London: Blackwell, 1999). The twelve authors and works (often excerpts thereof) include the following: René Descartes, *Meditations on a First Philosophy* (1641); John Locke, *Essay Concerning Human Understanding* (1690); David Hume, *An Enquiry Concerning Human Understanding* (1748); Gottfried Wilhelm Leibniz, *New Essays on Human Understanding* (1764); Immanuel Kant, *Critique of Pure Reason* (1781); John Stuart Mill, *A System of Logic* (1843); Friedrich Nietzsche, "On Truth and Lying in an Extra-Moral Sense" (1873); Charles Sanders Peirce, "The Fixation of Belief" (1877); Edmund Husserl, *The Idea of Phenomenology* (1907); Bertrand Russell, "Knowledge by Acquaintance and Knowledge by Description" (1912); Moritz Schlick, "On the Foundations of Knowledge" (1934); Ludwig Wittgenstein, *On Certainty* (1950).

Seeded Topic Modeling

Our first procedure was a variant of topic modeling that had proved illuminating in previous investigations and that involves the inclusion of a highly distinctive document in a larger, otherwise more homogenous corpus in order to "seed" the generation of topics characteristic of the added document.[10] In this case the collection of epistemology texts served as the seed document, functioning in essence as an additional, epistemological "novel" in the corpus. After adding the epistemology document to the other novels, we topic modeled the entire collection using the Mallet toolkit, experimenting with a variety of parameters to produce the highest proportion of coherent topics as well as to maximize the participation of the epistemology document in a single topic.[11] These aims were best achieved with a 50-topic run on corpora that had been divided into 500-word chunks and subject to a lengthy and carefully curated list of stopwords.[12] Table 1 indicates the "epistemology" topics for the English and German corpora—determined on the basis of the topic distributions for the epistemology document—along with the top 20 most frequently occurring terms in each.

Topic	Epist Part.	Top Words
Eng 32	63.10%	mind reason things human truth knowledge true thing religion good natural life general make part virtue power opinion sense ideas
Ger 12	65.50%	weiß wissen wahrheit gibt glauben dinge gar weise seele wesen darin denken vielleicht mensch glaube wahr grund geben frage wirklich

Table 1. Top 20 words in each epistemology topic

As the table makes clear, the epistemology document is overwhelmingly associated with topic 32 in English and topic 12 in German. The next highest ranking

---

[10]This essay assumes a basic familiarity with topic modeling and LDA. Readers who would like additional background, including information on parameter selection, may want to look at the following essays: David Blei, "Introduction to Probabilistic Topic Models,"; Matt Erlin,"The Location of Literary History: Topic Modeling, Network Analysis, and the German Novel, 1731-1864," *Distant Readings: Topologies of German Culture in the Long Nineteenth Century,* ed. Matt Erlin and Lynne Tatlock (New York: Camden House, 2014), 55-90; and Allen Beye Riddell, "How to Read 22,198 Journal Articles: Studying the History of German Studies with Topic Models," also in *Distant Readings*, 91-113.

[11]This optimization process is subjective but not arbitrary. Too many topics tend to generate word lists that reveal significant overlap, i.e. multiple topics where one has the sense that there should only be one, whereas too few leads to topics with a lack of thematic coherence. In our case the effort to maximize the participation of the epistemology document in a single topic simplified the process.

[12]The stopword list is available on Dataverse. We also eliminated terms occurring in less than 2% of the documents. Mallet parameters were set as follows: alpha = 50, optimize-interval = 10 and burn-in = 10.

topics are topic 14 (4.9%) in English and topics 0 and 47 (both 2.1%) in German - a distant second. In terms of proof of concept, it is reassuring to find a degree of semantic convergence between the epistemology topics in the two languages, including synonymous terms (e.g. reason/grund, things/thing/dinge, human/mensch, truth/wahrheit, knowledge/wissen, true/wahr), and equally reassuring to see that the constellations of terms are in line with what we would expect to find in works of epistemology (truth, reason, knowledge). [13]It should also be noted, however, that as an unsupervised method of classification, topic modeling is agnostic as to the question of what constitutes epistemological vocabulary. The algorithms simply determine which terms are most likely to co-occur in any given document as measured across a finite range of documents.

One could certainly attempt to identify epistemological content by other means, using a dictionary-based approach, for example, or a most-distinctive word test compared to some random corpus. Two primary advantages of topic modeling are 1) its relational character and 2) the fact that it produces results in the form of probability distributions that are easy to sort and compare. By topic modeling an entire corpus, we acquire a list of terms that are distinctive for the epistemology document as well as a measure of how frequently those terms occur in that document and in the various novels, and this measure can serve as a proxy for ranking the epistemological orientation of those novels. In this sense the method is similar to any dictionary-based approach. But we also generate additional semantic fields that are constitutive of the other texts in the corpus, enabling us to take a multi-layered approach to the question of epistemology. As will become apparent later in the essay, these other fields can provide a basis for determining which thematic preoccupations tend to correlate with a high percentage of epistemology terms, allowing for a more nuanced understanding of historical specificity.[14] These advantages notwithstanding, topic modeling results are, as a number of scholars have pointed out, sensitive to changes in the parameters of the analysis, and claims made on the basis of one particular set of results should always be understood as provisional.

Once we had established an epistemology topic for each corpus, we were then able to return to the novels, and, based on the percentage participation of any given novel in topic 12 or 32, generate lists of the top-ranking "epistemological"

---

[13]More unexpected is the range of other topics that prove to be nearly identical across national traditions, suggesting that English and German language novels in the eighteenth and nineteenth centuries have much in common in terms of broad thematic preoccupations. This similarity deserves additional investigation.

[14]For an additional discussion and justification of the "seeding" approach, see the appendix. Readers can use the topic modeling interface created by Stephen Pentecost of the Humanities Digital Workshop at Washington University to re-create our results and compare these with the results generated under different assumptions.

novels in each national tradition. The top 20 novels from each corpus are listed in Table 2. The thresholds for the top quartiles in the English and German corpora are 2.18% and 2.15%, respectively.

English Topic 32	%	German Topic 12	%
1756_amory_john_buncle	17.40%	1792_knigge_joseph_wurmbrand	10.45%
1809_more_coelebs_wife	15.80%	1821_goethe_meister_wanderjahre	8.04%
1815_peacock_headlong_hall	14.20%	1806_unger_bekenntnisse	7.94%
1759_johnson_rasselas	9.90%	1792_jacobi_aus_papieren	7.45%
1818_peacock_nightmare_abbey	8.00%	1802_novalis_lehrlinge	6.71%
1872_1901_butler_erewhon	8.00%	1835_gutzkow_wally	6.34%
1888_bellamy_looking_backward	7.90%	1916_reventlow_geldkomplex	5.31%
1726_swift_travels	5.90%	1915_rmueller_tropen	5.04%
1796_hays_emma_courtney	5.70%	1787_heinse_ardinghello	4.97%
1759_61_62_65_67_sterne_tristram_shandy	5.60%	1791_knigge_noldmanns_geschichte	4.86%
1792_holcroft_anna_at_ives	5.00%	1932_hofmannsthal_andreas	4.57%
1885_melville_billy_budd	4.90%	1919_sack_namenloser	4.45%
1749_fielding_tom_jones	4.90%	1795-6_tieck_william_lovell	4.39%
1743_fielding_journey	4.00%	1799_schlegel_lucinde	4.24%
1892_gissing_born_in_exile	3.90%	1805_klingemann_nachtwachen	4.15%
1794_godwin_caleb_williams	3.90%	1809_goethe_wahlverwandtschaften	4.13%
1806_edgeworth_leonora	3.60%	1917_sack_verb_student	4.09%
1749_fielding_governess	3.60%	1925_kafka_prozess	3.74%
1788_wollstonecraft_mary	3.60%	1793-4_hippel_kreuz_querzuege	3.74%
1798_brown_wieland	3.50%	1913_scheerbart_lesabendio	3.67%

Table 2. Top 20 novels in each corpus

As the word lists in Table 1 demonstrate, the semantic fields generated by the topic modeling algorithms often - though by no means always - feature nouns, with the consequence that topic modeling proves particularly effective at identifying overt thematic content. In order to gain a sense of just how effective, we can turn from the novels taken as a whole to the 500-word chunks that served as the basis for the analysis. The following brief excerpts from high-ranking passages in the English-language corpus provide examples of one of the more frequently occurring modes of epistemological engagement in the novels: that of an elucidation of the scope and limits of types of human knowledge. Words belonging to topic 32 have been highlighted.

The **difference** between **reason** and revelation seems to be the same as between the **eye** and the **light**; the one is the organ of vision, the other the source of illumination." "Take **notice**, Stanley," **observed** Mr. Tyrrel, "that if I can help it, I'll never **attend** your accomplished clergyman." "I have not yet completed the circle of his accomplishments," said Mr. Stanley, **smiling**; "besides what we **call book** learning, there is another species of **knowledge** in which some truly **good** men are sadly deficient: I mean an **acquaintance** with **human** nature. The **knowledge** of the world, and of him who **made** it; the **study** of the heart of man, and of him who has the **hearts** of all men in his hand, enable a minister to excel in the **art** of instruction; one **kind** of **knowledge** reflecting **light** upon the other.

Hannah Moore, *Coelebs in Search of a Wife* (1809)

Good and **evil** exist only as they are **perceived**. I cannot therefore **understand**, how that which a man perceives to be **good** can be in reality an **evil** to him: indeed, the **word** reality only signifies **strong** belief. Mr Escot. The views of such a man I contend are **false**. If he could be **made** to see the **truth**—— Mr Jenkison. He sees his own **truth**. **Truth** is that which a man troweth. Where there is no man there is no **truth**. Thus the **truth** of one is not the **truth** of another. Mr Foster. I am **aware** of the etymology; but I contend that there is an universal and immutable **truth**, deducible from the nature of **things**.

Thomas Love Peacock, *Headlong Hall* (1815)

With unreason the **case** is different. She is the **natural** complement of **reason**, without whose **existence reason** itself were nonexistent. If, then, **reason** would be non-existent were there no such **thing** as unreason, **surely** it follows that the more unreason there is, the more **reason** there must be also? Hence the **necessity** for the development of unreason, even in the interests of **reason** herself.

Samuel Butler, *Erewhon* (1872)

For, look at it in one way, all actions men **put** a bit of **thought** into are **ideas** - say, sowing seed, or **making** a canoe, or baking clay; and such **ideas** as these **work** themselves into **life** and go on **growing** with it, but they can't go apart from the material that **set** them to **work** and **makes** a medium for them. It's the nature of wood and **stone** yielding to the knife that raises the **idea** of shaping them, and with plenty of wood and **stone** the shaping will go on. I look at it, that such **ideas** as are mixed **straight** away with all the other elements of **life** are powerful along with 'em. The slower the mixing, the less **power** they have.

George Elliot, *Daniel Deronda* (1876)

"X—, however singular a **study** to some, is yet **human**, and **knowledge** of the world assuredly implies the **knowledge** of **human** nature, and in most of its varieties." "Yes, but a superficial **knowledge** of it, serving **ordinary** purposes. But for anything deeper, I am not certain whether to know the world and to know **human** nature be not two distinct branches of **knowledge**, which while they may coexist in the same heart, yet either may exist with little or nothing of the other. **Nay**, in an average man of the world, his constant rubbing with it blunts that **fine** spiritual insight indispensable to the **understanding** of the essential in certain exceptional characters, whether **evil** ones or **good**."

Herman Melville, *Billy Budd* (1885/1924)

Look at that basket, he said. —I see it, said Lynch. —In **order** to see that basket, said Stephen, your **mind** first of all separates the basket from the **rest** of the

visible universe which is not the basket. The first phase of apprehension is a bounding **line drawn** about the **object** to be apprehended. An aesthetic image is **presented** to us either in **space** or in **time**. What is audible is **presented** in **time**, what is **visible** is **presented** in **space**. But, temporal or spatial, the aesthetic image is first luminously apprehended as selfbounded and selfcontained upon the immeasurable background of **space** or **time** which is not it. You apprehended it as ONE **thing**. You see it as one whole. You apprehend its wholeness. That is INTEGRITAS. —Bull's **eye**! said Lynch, laughing.

James Joyce, *A Portrait of the Artist as a Young Man* (1916)

Read against the backdrop of more structural approaches to the epistemological engagements of the novel, these passages remind us that novels also regularly address epistemological questions head on and in some detail. It is no doubt true, as John Bender argues, that the "*implicit* ambitions of the new novel" [my italics] in the eighteenth century positioned the genre within the same discursive network as the experimental practices of the natural sciences. It is equally true, however, that the novel frequently participates in debates "about method and the nature of knowledge" at the level of *explicit* content, often through dialogue, and that this level is arguably as significant as any other in terms of the cultural work done by the genre.[15]

As revealing as these passage are, however, they also raise the question of what the topic models are missing. With regard to *Billy Budd*, for example, even though the algorithm correctly identifies the epistemological orientation of the passage, it fails to classify number of terms (e.g. "insight") that can certainly be characterized as epistemologically relevant. Moreover, the passage demonstrates how epistemological relevance can find expression in linguistic features less conspicuous than an explicitly philosophical vocabulary. In the case of the *Billy Budd* excerpt, the epistemological moment is strengthened by a particular narrative perspective. More precisely, the narrator not only discusses branches of human knowledge; he also adopts a stance toward the certainty of his own knowledge of that knowledge, through such terms as "assuredly" and phrases as "I am not certain." Such terms might be understood as marking a second order epistemological engagement, one that bears similarities to what Andrew Piper, in his recent essay on fictionality, refers to as the "access to the knowledge of knowing" that is typical of the novel as genre.[16] Our second method attempted to address this less conspicuous mode of epistemological engagement.

---

[15] Bender, "Novel Knowledge," 290.

[16] Andrew Piper, "Fictionality," *Journal of Cultural Analytics*. Accessed 12/20/16.

Epistemic Modals

Passages such as the one from Melville involve a subjective positioning on the part of characters or the narrator vis-à-vis the content of their utterances. In other words, they involve questions of modality, a category of meaning that has recently been the subject of extensive analysis on the part of the linguists and philosophers of language, and that pertains to words and phrases used to communicate relations of probability and necessity.[17] Within the general category of modality, moreover, scholars distinguish between variants, depending on whether the relations of necessity and possibility pertain to knowledge about the world (epistemic), to duty (deontic), to desires (bouletic), or to circumstances (circumstantial). Epistemic modality is most relevant in the current context, and especially the use of epistemic modal adverbs and adjectives, which, in the words of Jan Nuyts "may be considered the 'purest' expressions for epistemic modality, in the sense that they are the most precise and specific means available for marking the degree of likelihood of a state of affairs."[18] In English and German, one can locate these terms on an epistemic scale that ranges from certainty to probability, possibility, doubt, and impossibility. Also of interest are utterances that rely on "epistemic indefinites," defined as "indefinite determiners or indefinite pronouns that signal ignorance on the part of the speaker" and thus convey "information about her epistemic state."[19] An example from German would be "irgend" or "irgendein" (some X or another), which indicates partial knowledge of a situation but also the absence of some information.

Our initial analysis using topic models led us to conclude that measuring the frequency with which modal expressions appear in the novels could offer a further perspective on the question of their epistemological moment, one that picks up on more subtle epistemological indicators and thus might either confirm or qualify the results of the topic modeling. We took a more targeted approach in this case. Rather than relying on unsupervised classification to identify constellations of terms, we instead pre-selected a list of 30 adverbs with epistemic resonance, building on the work of specialists in the field. We then charted their frequency across time, both for the novels as a whole and for shorter, in this case, 1,000-word chunks.[20]

---

[17] Kai von Fintel, "Modality and Language," *Encyclopedia of Philosophy—Second Edition*, ed. Donald M. Borchert (Detroit: MacMillan Reference USA, 2006

[18] Jan Nuyts, *Epistemic Modality, Language, and Conceptualization* (Amsterdam: John Benjamin's Publishing Company, 2001), 56.

[19] Luis Alonso-Ovalle and Paula Menéndez-Benito, "Epistemic indefinites: An overview," *Epistemic Indefinites: Exploring Modality Beyond the Verbal Domain,* ed. Luis Alonso-Ovalle and Paula Menéndez-Benito (Oxford: Oxford UP, 2014), 2.

[20] For reasons that will become clear in the course of the discussion, this method was applied only

Aggregates and Outliers

The results of the topic modeling discussed above provide a useful starting point for thinking about the epistemologies of the novel in historical context, especially as regards distinctions within and between the individual corpora. Returning to Table 2, we see that in both lists, the eighteenth and early nineteenth centuries are significantly overrepresented with regard to the corpus as a whole, a result that aligns with those scholarly narratives that posit the rise of the novel in the eighteenth century as an epistemological project. On the other hand, the English corpus includes four novels in the top twenty that were written or published in the second half of the nineteenth century (1872, 1885, 1888, 1892), whereas the German corpus has none. Is it possible that German realist novels have, on aggregate, fewer epistemological preoccupations than their English counterparts? One way to get a better sense of general shifts in topic participation as they relate to our existing understanding of literary history is to consider average epistemology topic percentages across each corpus for separate periods, as I have done in Table 3.

Period	English	Topic 32	Δ	German	Topic 12	Δ
Enlightenment/Neoclassical (1660-1790)	42	2.69%		33	1.80%	
Romantic (1791-1830)	47	2.71%	0.51%	65	2.11%	17.22%
Nineteenth-Century/Victorian (1831-1901)	132	1.51%	-43.87%	108	1.46%	-30.81%
Modern (1902-)	41	1.13%	-25.45%	55	1.73%	18.49%
Total Novels/Average for Corpus	262	1.86%		261	1.72%	

Table 3. Epistemology topic averages by period.

The literary-historical periods used in this table were chosen with an eye to cross-cultural validity as well as to making them as non-controversial as possible in light of the small size of the two corpora. Comparisons of actual percentages across the two national traditions carry no meaning, because each corpus was topic modeled separately.[21] What is comparable, however, are general trends in the rise and fall of the topic percentages, and in this regard the two sets of novels prove rather similar, with one notable exception. In both cases one finds a high-point in the romantic period, followed by decline in the nineteenth century. The

---

to the German corpus.

[21] One can, on the other hand, isolate the twenty-eight American novels in the English-language corpus that appeared between 1831 and 1901and establish that their aggregate participation in the epistemology topics is significantly higher than that of the British novels published in the same period: 1.90% versus 1.41%.

major discrepancy between the two corpora pertains to the twentieth-century. The percentage participation of the German novels from that period shifts upward (18.49%), whereas that of the English novels declines (- 25.46%).

Another view of the data allows us to approach these shifts from a more granular perspective. Figures 1 and 2 plot the epistemology topic percentages of the novels in the English and German corpora published between 1791 and 1901, with year of publication and percentage participation in the epistemology topics as the respective x and y axes.[22]
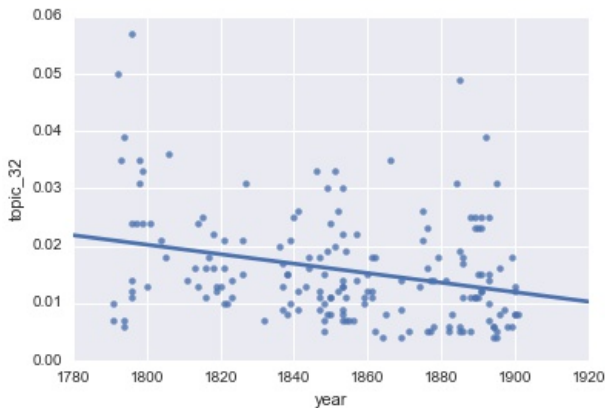

Figure 1. Scatter plot of English novels 1791-1901 with regression line (R2=.081, F(1,172)=15.13, p<.001)

---

[22] In order to ensure that the regression lines were not being disproportionately affected by a small number of very high-ranking texts, we eliminated any texts with a topic 32 or topic 12 percentage of more than 2.5 standard deviations above the mean. This filter removed 6 novels from the German and 5 novels from the English corpus.
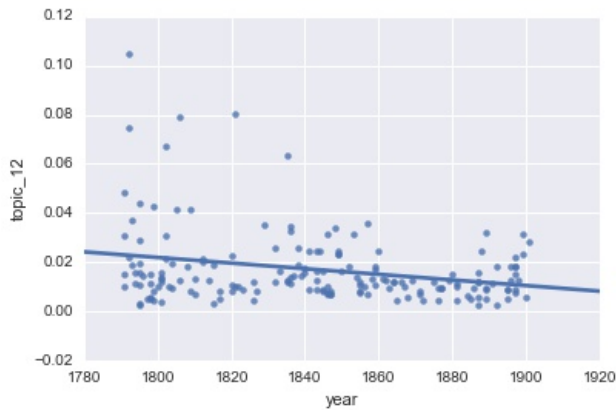
Figure 2. Scatter plot of German novels 1791-1901 with regression line (R2=.034, F(1,165)=5.786, p=.017)

The two plots indicate that the aggregate drop in the means of the two corpora actually corresponds to a downward trend. These are noisy patterns, to be sure, but they align well with other recent computational work on the nineteenth-century novel. To my mind, the most plausible hypothesis for the decline is a variant of the argument presented by Ryan Heuser and Long Le-Khac in their Stanford Literary Lab pamphlet from May 2012. Working with an English-language corpus of over 2,500 novels and a series of semantic taxonomies they constructed using the OED's historical thesaurus, Heuser and Le-Khac present compelling evidence of a significant drop in the frequency of terms denoting abstract values. This drop, moreover, proved to be historically correlated with an increase in what they call "hard seed" terms - a collection of more concrete and physical description words.[23]

Considering the overwhelming predominance of abstract terms in topics 32 and 12, the aggregate data from the epistemological topics can serve as a corroboration of these results, one indicating that the decline in abstract values holds for both German and English literature, at least until the beginning of the twentieth century. Further support for this hypothesis can be found in categories identified by the topic models that correspond quite closely to Heuser and Le-Khac's "hard seed" field. In addition to the epistemological topics, in other words, the topic modeling also generated parallel topics across corpora in which concrete terms predominate. Most notable among these are a "body parts" topic (G30: hand,

---

[23] Ryan and Long Le-Khac, "A Quantitative Literary History of 2,958 Nineteenth-Century British Novels: The Semantic Cohort Method," *Stanford Literary Lab Pamphlet 4*, 27.

augen, gesicht, kopf, blick, stand; E48: eyes, hair, face, black, head, look) and a "domestic interiors" topic (G39: tür, zimmer, fenster, ging, stand; E30: door, room, house, window, open) The regression lines for both of these topics show a significant increase over the same period.[24]

We can thus link the aggregate results of the topic modeling to a broader decline over the course of the nineteenth century in novelistic abstraction and a correlated rise of concrete terms. By honing in on the question of epistemology, however, we are able to add some additional nuance to the picture presented by Heuser and Le-Khac. First and foremost, the aggregate trends should not blind us to the fact that the decline is not constant but ebbs and flows, at least as far as the epistemological preoccupations identified by the topic models are concerned. In fact, the outliers may ultimately be more productive as a starting point for literary-historical arguments than aggregate trends. Novels with significant (e.g. top decile) epistemological engagements occur across the historical span of both corpora, suggesting that there may be value in positing the existence of an "epistemological novel" as a hitherto neglected generic category along the lines of the "conversional novel" recently proposed by Andrew Piper. Indeed, inasmuch as these high-scoring novels feature explicit engagements with questions of knowledge and cognition rather than affect, we can perhaps see it as a kind of companion genre.[25]

The Epistemology of What?

While the topic percentages can give us a sense of which novels share a semantic field with the epistemology document, they tell us nothing about the how the terms in this field are deployed in concrete instances. As the excerpts cited previously suggest, and as a review of other top-ranking passages confirms, novels do incorporate explicit, abstract reflections on the sources and validity of knowledge. But they also offer a range of examples of applied, and often embedded, epistemological inquiry. I will have more to say about embedding—the "how" of novelistic epistemologies—shortly. Before turning to that subject, however, we can use the results of the topic modeling to identify some distinctions at the level of content—the "what"—and thus begin to develop a preliminary typology that speaks to national particularities as well as cross-cultural parallels.

---

[24]In their discussion of the hard seed words, Heuser and Le-Khac refer specifically to terms that designate body parts as well as physical qualities.

[25]Andrew Piper, "Novel Devotions: Conversional Reading, Computational Modeling, and the Modern Novel," *New Literary History* 46.1 (2015): 63-98.

At issue is the question of what other topics predominate in those novels that rank highly for epistemological content. One way to answer this question is to select a subset of those novels - we took the top 25 - and then determine the top non-epistemology topics for each - we chose the top three. We then aggregated the results to establish which of the non-epistemology topics most frequently rank among the top three non-epistemology topics for novels in our selection. Table 4 lists the results and includes the number of novels and their titles together with the top 20 words for each topic. It also includes an interpretive label for each topic for the purposes of orientation, but it should be noted that the topic words are disparate enough that there is room for debate here, especially in the case of the German topics.[26]

Topic	Label	Novels	Top 20 Terms	Titles
7	life in society	13	time made house friend person found gentleman company received immediately good fortune account manner gave told part make present	1726_swift_travels; 1742_fielding_joseph_andrews; 1743_fielding_journey; 1749_fielding_governess; 1749_fielding_tom_jones; 1756_amory_john_buncle; 1759_johnson_rasselas; 1780_holcroft_alwyn; 1793_dibdin_younger_brother; 1792_holcroft_anna_st_ives; 1794_godwin_caleb_williams; 1815_peacock_headlong_hall; 1818_peacock_nightmare_abbey
33	love and af-fection	8	heart love happiness life mind friend happy affection dear passion pleasure tears hope moment tender soul felt feelings long	1759_johnson_rasselas; 1749_fielding_governess; 1780_holcroft_alwyn; 1788_wollstonecraft_mary; 1796_hays_emma_courtney; 1806_edgeworth_leonora; 1809_more_coelebs_wife; 1818_peacock_nightmare_abbey
14	mental ac-tivities and objects	6	life felt made mind thought sense time feeling fact things knew long moment found position conscious sort point part	1866_eliot_felix_holt; 1872_1901_butler_erewhon; 1885_melville_billy_budd; 1888_bellamy_looking_backward; 1892_gissing_exile; 1903_butler_ernest_pontifex
Topic	Label	Novels	Top 20 Terms	Titles
0	the human spirit	11	leben geist zeit seele mensch kraft lebens kunst gedanken wer freiheit gibt großen wesen nie wahrheit menschheit dichter liegt	1787_heinse_ardinghello; 1792_knigge_wurmbrand; 1795_6_tieck_lovell; 1799_schlegel_lucinde; 1802_novalis_lehrlinge; 1805_klingemann_nachtwachen; 1806_unger_bekenntnisse; 1835_gutzkow_wally; 1917_sack_student; 1919_sack_namenloser 1932_hofmannsthal_andreas
6	abstract unclas-sified	7	großen lassen zeit genug wenig allein sache geschichte teil weniger guten mittel geben tugend weise endlich vielleicht große wenigstens	1773-6_nicolai_nothanker; 1781_wieland_abderiten; 1787_heinse_ardinghello; 1791_knigge_noldmann; 1792_knigge_wurmbrand; 1806_unger_bekenntnisse 1821_goethe_wanderjahre
11	abstract unclas-sified	6	zeit leben schien fühlte nie ließ weise fand wusste tage seit gesellschaft ersten gab blieb allein wesen wenig familie	1799_schlegel_lucinde; 1806_unger_bekenntnisse; 1809_goethe_wahlverwandtschaften; 1821_goethe_wanderjahre; 1829_fouque_resignation; 1857_stifter_nachsommer
24	abstract unclas-sified	6	lassen vielleicht hand weiß geben zeit mann sache wissen freund bitte möglich rief wort kommen finden ehre nehmen glauben	1781_wieland_abderiten; 1792_knigge_wurmbrand; 1809_goethe_wahlverwandtschaften; 1821_goethe_wanderjahre; 1829_fouque_resignation; 1926_kafka_schloss

Table 4. Topics that track with epistemology

These lists prove revealing on a number of levels. Before turning to the lists themselves, however, it is important to note that epistemological content is highly ecumenical in its affiliations, that is to say, it appears together with a wide range

---

[26]German topics 6, 11, and 24 have simply been labelled "abstract unclassified" for the reason that although they include a preponderance of abstract terms, they also lack any obvious thematic coherence. An analysis of high-ranking passages from the novels could provide additional guidance here, but establishing a definitive label for these topics is not crucial to my argument.

of other thematic concerns. The tables include only the non-epistemological topics that ranked among the top three in the highest number of novels, but in fact no less than 21 topics from the German corpus and 24 topics from the English ranked among the top three non-epistemological topics for at least one high-epistemology novel. Nonetheless, it is also the case that for both sets of novels, the epistemology topics occur most frequently in conjunction with other abstract topics. It appears, in other words, that epistemological engagement tends to manifest as part of a broader "philosophical" orientation or at least an emphasis on interior states. This is perhaps not surprising, but a careful consideration of the topic words also suggests a few notable divergences between the national literary traditions. Limiting ourselves to the most frequently occuring non-epistemology topic for each corpus, we find a suggestion of life in society in English topic 7 as compared to the more abstractly intellectual concerns of German topic 0, despite the fact that in both cases the majority of novels come from the eighteenth and early nineteenth centuries. Things become somewhat more complicated when we bring in additional topics, since English topic 14 is also highly abstract, albeit less in the sense of broad categories of intellectual inquiry (in the German corpus: freedom, truth, art, humanity, history, society) than in the sense of terms denoting individual perceptual faculties (felt, thought, knew, sense) as well as terms related to mental objects and attitudes (fact, point, position). It is also associated with a group of later nineteenth-century novels that place heavy emphasis on criticism of societal institutions, either in direct form or *ex negativo* through the presentation of utopian communities.

Even as they demonstrate a range of similar concerns across time periods and across both corpora, then, these results also point to differences in emphasis between the German and English novels. Since 1810 at the latest, when Germaine de Staël famously referred to Germany as the "Land der Dichter und Denker" (land of poets and thinkers), German literature has been seen to possess a uniquely cerebral, philosophical, or interior quality. While one finds no evidence that German novels are more epistemological than English novels in any straightforward sense, except perhaps at the beginning of the twentieth century, they do appear to be *differently* epistemological, and the predilection for a certain mode of generalization and abstraction around 1800 may help to explain the stereotype.

A further perspective on the non-epistemological content of the epistemological novels can be acquired by approaching the question from the opposite direction, that is to say, by identifying other topics that predominate in the novels that rank lowest for epistemological content. Table 5 includes the single highest-ranked non-epistemology topic that occurs in a subset of the bottom 20 episte-

mology novels for each corpus (participation levels in the epistemology topics from roughly .2% to .5%).

Topic	Novels	Top Words	Titles
30	9	hand augen gesicht kopf blick stand stimme hände trat ging lippen legte zurück arm leise fragte hielt ließ zog	1872_80_freytag_ahnen; 1887_sudermann_sorge; 1892_christen_mutter; 1895_ganghofer_hubertus 1907_keyserling_dumala; 1911_keyserling_wellen 1914_keyserling_häuser; 1914_ganghofer_ochsenkrieg 1918_ball_flammetti
6	8	hand face eyes head back hands looked arm stood cried moment arms turned round lips held put suddenly voice	1882_stevenson_treasure_island; 1894_freeman_pembroke; 1895_conrad_almayers_folly; 1905_orczy_pimpernel; 1903_london_call_wild; 1911_wharton_ethan_frome; 1913_lawrence_sons_lovers; 1930_mansfield_aloe

Table 5. Topics that are high in the low-epistemology novels

The overlap between the two corpora in this context is quite striking, not only in terms of historical distribution (very late nineteenth and early twentieth century) but also in terms of the top words: eyes/augen, face/gesicht, head/kopf, stood/stand, voice/stimme, back/zurück. The combination of body parts and verbs (especially together with "suddenly" - the German equivalent "plötzlich" is actually the next word in the German list) suggests moments of encounter, of movement, in short, of plot rather than reflection, and the inclusion of such action-packed popular novels as *Treasure Island* and *The Call of the Wild* reinforces this impression. One can interpret these results as offering further support of Heuser's and Le Khac's argument about the decline of abstraction - the very lowest epistemology quotients are associated with a high level of concrete descriptive vocabulary and publication dates in the late nineteenth- and early twentieth-century - even as they add nuance to that argument by demonstrating the ongoing existence of clusters of novels that buck the trend. The existence of such clusters, moreover, together with the repeated appearance of epistemological outliers throughout the century, point to the possibility of linking the decline that appears in the aggregate data to an expansion and segmentation of the market. The nineteenth century bears witness to a dramatic expansion in the size of the literary market, and especially in the market for popular fiction. If mass market novels in general tend to include more plot and less reflection, then the expansion of the market will lead to a great proportion of such novels being published and to a corresponding decline in the relative degree of abstraction, even if not all works participate in the trend. The question of participation is not simply one of canonical versus popular works or authors - Joseph Conrad and D.H. Lawrence rank low for both of the abstract topics discussed thus far (32 and 14). Genre may be a better category for trying to understand the fate of abstraction, and in this regard the high epistemology novels associated with topic 14 prove especially interesting, because they also challenge us to rethink our generic categories. Butler's *Erewhon* and Bellamy's *Looking Backward* are utopian novels, but

what about Eliot's *Felix Holt,* Melville's *Billy Budd*, Gissing's *Born in Exile,* and Butler's *Ernest Pontifex* (*The Way of All Flesh*). Might we tie these together on the basis of their concern with social criticism? With human and societal evolution? With a particular narrative perspective?

The final possibility reminds us of the previously mentioned shortcoming of the topic modeling approach; namely, that it can bind us to thematic content in a manner that de-emphasizes the multiple levels at which epistemological engagement can play out. In fact, to the extent that one can speak of a uniquely novelistic mode of such engagement, one would expect it to be not merely applied, in the previously mentioned sense that it departs from the narrowly disquisitional approach of philosophical texts, but also embedded. Novels that explicitly address "the epistemology of x" at the level of content may not even be the most epistemological novels in a general sense. Our topic-modeling derived measures provide a useful proxy for identifying general trends, but any method according to which Woolf's *To the Lighthouse* ranks well below average (1.3%) on the epistemology scale is not telling the full story.

Finding a proxy for sub- or supra-lexical features of epistemological engagement is no easy task. In the remainder of this essay, I will describe some first steps taken in this direction, which are limited to the German novels and based on the epistemic modal adverbs discussed previously. Our first effort in this regard was inspired by results of the initial topic modeling. For the German corpus, the second highest ranking topic for the epistemology document (47), although far less significant in terms of percentage participation (2.1% versus 65.5%), caught our attention because it included, somewhat atypically, more adjectives and adverbs than nouns, and because it was dominated by the novels of Franz Kafka. Two of Kafka's novels, famous for narrating failed quests for absolute knowledge, had ranked highly for the primary epistemology topic, but the participation levels for this second topic made it clear that topic 47 was more or less a Kafka topic. It was at this point that we began investigating the scholarship on epistemic modality and decided to re-run the topic modeling on a corpus that had been part-of-speech tagged and then reduced to include only adverbs and adjectives. Dividing the new adjective/adverb "texts" into 100 word chunks and using the same stopword parameters as before, we set the number of topics at 15, which generated two topics that were the most characteristic of the "dehydrated" epistemology document. Interestingly, these topics split along two significant lexical axes. Topic 12, the second highest ranking topic, focused on relationships of magnitude (groß, große, weniger, größten) as well as between part and whole (darin, besonders, allgemeinen, allgemeine, verschiedenen) and the novels that shared high participation in this topic tended to come from earlier periods.

Topic 14, on the other hand, was both skewed toward the twentieth century and was overwhelmingly populated with epistemic adverbs (vielleicht, wirklich, gar, sogar, möglich, gerade, natürlich, erst, allerdings, irgend, eigentlich, wahrscheinlich, deshalb, genau, ebenso).[27] Table 6 lists the top 20 novels for this "epistemic adverbs" topic and includes the epistemology document for comparison. The question that emerged from these results was whether a case could be made that one mode of epistemological engagement is more frequently combined with or even displaced by another as the nineteenth century progresses, that is to say whether we could identify a historical shift away from explicit epistemological reflection and toward epistemic narration, understood as a narrative perspective that foregrounds the epistemic relationship between the speaker or narrator and the content of his or her utterance.[28]

1925_kafka_prozess	27.80%
1926_kafka_schloss	26.90%
adj_adv_de_epistemology	25.90%
1916_reventlow_geldkomplex	25.00%
1899_may_am_jenseits	24.80%
1927_kafka_amerika	23.50%
1904_may_friede	23.40%
1902_may_silb_löwen_III	22.80%
1903_may_silb_löwen_IV	22.30%
1910_may_winnetou_IV	22.30%
1913_reventlow_hrn_dames	20.50%
1908_schnitzler_weg_ins_freie	18.80%
1910_rilke_malte	18.20%
1915_meyrink_golem	18.20%
1889_conradi_adam_mensch	18.10%
1912_reventlow_paul_pedro	18.00%
1925_reventlow_selbstmordverein	17.90%
1919_essig_taifun	16.70%
1859_ruppius_vermächtnis	16.60%
1897_spielhagen_zeitvertreib	16.50%
1928_schnitzler_therese	15.50%

Table 6. Top ranking novels for the epistemic adverbs topic

From a literary-historical perspective, the most striking result of this set of results is the fact that the canonical modernist Franz Kafka appears together at the top of the list with Karl May, the popular and prolific author of adventure novels of the American West (and, occasionally, the Far East). A quick review of the MLA and the *Bibliographie der deutschen Sprache und Literatur* (BDSL) databases reveals a total of two articles that consider them in tandem.[29] Given the intriguing possibility that there may be an uncharted influence or at least an

---

[27] (large, large, less, largest); (therein, particularly, general, general, different); (possibly, really, even, even, possible, precisely, actually, probably, therefore, exactly, likewise)

[28] See Piper, "Fictionality," Accessed 12/20/16.

[29] Sebastian Susteck, "Die überstimmte Fremde: Karl Shatterhand, Karl Rossmann, Rhetorik, *Wirkendes Wort* 56.1 (2006): 15-32. Ulf Abraham, "Die Angst vor der Entdeckung und die Entdeckung der Angst: ein Motiv bei Franz Kafka und Karl May," *Deutsche Vierteljahrsschrift für Literaturwissenschaft und Geistesgeschichte* 59 (1985): 313-40.

unrecognized stylistic commonality in play here, a comparison of some of the reconstituted passages of these two authors provides a good starting point for further investigation of the epistemic hypothesis.[30] What one finds in both sets of passages is a remarkably similar effort on the part of protagonists to manage epistemological uncertainty in unfamiliar environments. Frequently this effort entails evaluating competing interpretations of specific events or behaviors, and even more frequently the protagonists are engaged with interlocutors whose reliability remains open to question or who are in need of instruction. In other words, dialogue figures prominently and narrators tend to be either first-person or covert. A typical example from May is the following, which comes from the travel narrative *Und Friede auf Erden* (1904):

> "Sie tat, als ob sie durch diese Mitteilungen beruhigt worden sei, war es aber wahrscheinlich nicht, wenigstens nicht ganz wie mir ja gerad durch Ihre Wortkargheit bewiesen wurde."

> She acted as though she had been reassured by this report, but she probably hadn't been, at least not entirely, as was proven to me by her very silence.

Here the uncertainty finds expression not only in the use of epistemic adverbs ("wahrscheinlich") and quasi-epistemic phrases ("nicht ganz") but also in the use of the subjunctive form ("tat, als ob . . . sei"). Also noteworthy here is the spectrum of possibility presented. The first-person narrator interprets the apparent relief of the woman as insincere but also indicates that his interpretation is only probably correct. One finds precisely these same elements in Kafka, as in the following brief exchange between K. and the baliff from *Der Prozess*:

> "Es würden vielleicht, fuhr K. fort, auch noch andere Ihrer Beamten und vielleicht sogar alle das gleiche verdienen. Ja ja, sagte der Gerichtsdiener, als handle es sich um etwas Selbstverständliches."

> It may in fact be the case, K. continued, that some of your other officials and perhaps even all of them earn the same. Right, right, said the baliff, as if it were a matter of something self-evident.

Again we have a two-tiered equivocation, in the sense that reaction of the baliff casts doubt onto the reasonableness of K's claim but is itself rendered ambigu-

---

[30] As previously explained, the topic modeling was run on "dehydrated" texts that included only adjectives and adverbs. These have to be reconstituted, and the reconstituted passages are not necessarily equal in length. In other words, in one novel a chunk of 100 adjectives and adverbs might correspond to a 500-word passage, whereas in another it might be a 1,000-word passage, in which the terms are dispersed across a wider range.

ous through the use of the subjunctive (*as if* it were a matter of something self-evident).

Both authors, then, are concerned with delineating what we might term a hermeneutics of intersubjectivity in situations of threatening opacity, except that in May this opacity arises from travel to exotic lands whereas in Kafka it has become a feature of an environment that ought to be familiar. One should also note that despite the stylistic and rhetorical similarities between the two authors, May documents both the opacity as well as many successful efforts to render it transparent, whereas in Kafka the hermeneutics ultimately seems to spin out of control in a kind of mise-en-abyme of interpretation. Most significant in the context of the current analysis, however, is the fact that these passages provide support for the notion of an evolution toward a greater emphasis on epistemic narration, that is to say, toward novels in which epistemological concerns have become more deeply embedded in the narrative structure. This is not to claim that these novels have no explicitly philosophical content. Five passages from May's novels and three from those of Kafka also appear in the list of the top 100 high-ranking passages for epistemology topic 12. The point, rather, is that this content represents only part of what makes the novels epistemologically interesting, and, even more importantly, that there may be other novels that lack such content but nonetheless exhibit a strong epistemological orientation at the level of narration.

Establishing or refuting the existence of a historical phenomenon of the sort under consideration here will require additional analysis and significance testing, especially if one wants to incorporate a cross-cultural comparison. Initial experiments with the tagged English corpus did not generate any topics that foregrounded epistemic modality to the same degree as the German corpus. Part of the challenge stems from the fact that adjectives and adverbs constitute only one of the linguistic elements through which modality finds expression. The previous discussion of the passages from Kafka and May alluded to mood, in this case the subjunctive, as a closely related or even subsidiary category and another means of indicating a particular epistemic or propositional stance. In the linguistic and philosophical literature on the topic, modal auxiliaries (must/should/might/may) and semimodal verbs (has to/ought to/needs to) tend to appear most frequently in illustrative examples.[31] It may be the case that modality is more likely to be expressed through such verbs or through some other means (nouns, conditionals) in English than in German. The advantage of focusing on adjectives and adverbs, however, is that their connection to specifically epistemic relations of probability and necessity is often more

---

[31] Kai von Fintel, "Modality and Language," Accessed 2/8/16.

straightforward than is the case with other terms.[32]

In an effort to corroborate the results we were seeing from the topic modeling, we undertook a further investigation involving a targeted approach to the question of epistemic stance. Rather than relying on the topic modeling algorithms to identify constellations of terms, we pre-selected a list of 30 adverbs with epistemic resonance and simply charted their aggregate frequency across time.[33] The line in figure 3 represents the 20-year rolling mean of the per-novel frequency of these adverbs, calculated in relation to the total number adjectives and adverbs appearing in the novel.
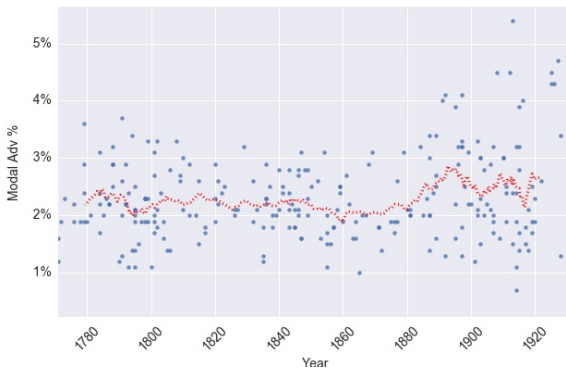

Figure 3. Modal adverbs and adjectives as % of total adverbs and adjectives.

Except for the steady increase in the frequency of these modal terms between roughly 1870 and 1890, the rolling mean offers little to support the notion of a shift. Indeed, to the extent that we are in fact dealing with a displacement of one mode of epistemological engagement by another, we would expect a negative correlation between the frequency of modal adverbs and participation in epistemology topic 12, but in fact the two variables have a weakly positive correlation

---

[32]For example, As Kai von Fintel explains, the English semimodal "have to" demonstrates a remarkable flexibility of modal meanings, not just epistemic ("It has to be raining") but also what is known as circumstantial ("I have to sneeze") and deontic ("Visitors have to leave by six pm"). See "Modality and Language."

[33]In German, adverbial forms are often distinguished from adjectival forms by the absence of an ending. For this analysis, we only considered these uninflected forms, some of which, however, may also function as predicate adjectives. The list of adverbs is as follows: vermutlich, voraussichtlich, tatsächlich, angeblich, natürlich, gewiß, offensichtlich, weitaus, sozusagen, keineswegs, keinesfalls, wohl, unbedingt, wahrscheinlich, wirklich, zweifellos, anscheinend, irgendwie, vielleicht, selbstverständlich, scheinbar, sicher, sicherlich, bestimmt, offenbar, kurzerhand, quasi, gewissermaßen, möglicherweise, eigentlich.

for all of the literary periods identified in Table 3 (Enlightenment: .11; Romanticism: .11; Nineteenth Century: .09; Modern: .30).

As before, however, it proves productive to redirect our attention away from the aggregate to the outliers. Of particular interest are novels for which the discrepancy between the two measures of epistemological engagement is the highest, together with the question whether they occur more frequently in a particular period. Calculating a z-score for each novel for each of the two measures enables us to compare them on the same scale. If we then subtract the topic 12 z-score from the modal adverb z-score, we can rank the novels from lowest (meaning a relatively higher topic 12 score) to highest (meaning a relatively higher modal adverb score).

category	novel	adv-z	topic-z	diff
Nineteenth	1847_gerstäcker_flusspiraten	1.029133	-0.641389	1.670522
Nineteenth	1868_felder_reich_und_arm	0.857038	-0.837833	1.694871
Nineteenth	1897_janitschek_amazonen_schlacht	1.14567	-0.592456	1.738126
Nineteenth	1887_fontane_irrungen_wirrungen	1.458693	-0.390339	1.849032
Nineteenth	1884_anzengruber_sternsteinhof	1.14838	-0.761241	1.909621
Romanticism	1799_vulpius_rinaldini	1.315055	-0.656991	1.972046
Enlighten	1788_ehrmann_ninas_briefe	1.167351	-0.827905	1.995256
Nineteenth	1889_conradi_adam_mensch	1.507476	-0.586073	2.093549
Modern	1928_schnitzler_therese	1.5007	-0.657701	2.158401
Modern	1907_keyserling_dumala	1.175481	-1.047042	2.222524
Nineteenth	1892_fontane_fr_jenny_treibel	2.399118	0.085523	2.313594
Modern	1912_reventlow_paul_pedro	2.930308	0.554293	2.376015
Modern	1915_boy_ed_vor_der_ehe	2.161979	-0.294599	2.456578
Nineteenth	1897-8_fontane_stechlin	2.476357	-0.135033	2.61139
Nineteenth	1891_raabe_stopfkuchen	2.259544	-0.414451	2.673995
Nineteenth	1895_fontane_effi_briest	2.18095	-0.592456	2.773406
Modern	1913_reventlow_hrn_dames	4.133618	1.067032	3.066586
Modern	1908_schnitzler_weg_ins_freie	3.000773	-0.246375	3.247147
Modern	1925_reventlow_selbstmordverein	2.724336	-0.660537	3.384874
Modern	1927_kafka_amerika	3.278564	-0.815848	4.094412

Table 7. 20 Novels with the highest discrepancy between epistemology topic and modal adverb frequencies

Of the 32 novels that have a modal adverb score relative to the topic 12 score that was more than one standard deviation above the mean, 24, or 75%, were published after 1865.[34] The top ten include six novels from the "modern" category: Kafka's *Amerika*, three works by the feminist author Fanny zu Reventlow (*Paul Pedro*, *Herrn Dames Aufzeichnungen*, and *Der Selbstmordverein*), one by the Austrian Arthur Schnitzler (*Der Weg ins Freie*), and one by the prolific popular novelist and *salonière* Ida Boy-Ed (*Vor der Ehe*). This list also includes four novels from the very late nineteenth century, three by Theodore Fontane (*Effi Briest, Frau Jenny Treibel,* and *Der Stechlin*) and one by Wilhelm Raabe (*Stopfkuchen*).

Even in the absence of confirmation from traditional measures of statistical

---

[34] By comparison, the corpus as a whole includes 104 of 261 novels published after 1865, roughly 40%.

significance, this group of novels raises the possibility of a cluster of very late nineteenth- and early twentieth-century works - a cluster that presumably also includes Karl May and others - in which a unique mode of epistemological engagement comes to the fore. One cannot speak of a radical shift even with regard to these works, given that a number of them also register these engagements more explicitly (other novels by Kafka and Reventlow rank high for the epistemology topic as well). But the results provide additional evidence of a recalibration in a certain subset of novels, a recalibration that could have real literary-historical significance.

Some preliminary textual evidence for this shift has already been discussed, and a consideration of key passages from Fontane and Raabe, the nineteenth-century authors in the group, can help to reinforce the case for additional research. In order to identify passages with high modal content, we simply broke the four novels into 1,000-word chunks and then calculated the 20 chunks with the highest cumulative number of modal adverbs from our list. This approach generated three passages from *Stopfkuchen*, four from *Frau Jenny Treibel,* five from *Effi Briest*, and eight from *Der Stechlin*. All but one of these passages (the exception is from *Effi Briest*) revolve around dialogue (as opposed to the voice of the narrator), and a strong case can be made for the claim that all of them are concerned with what I previously termed managing epistemological uncertainty. In these texts, however, questions that arise about the behavior of others, or about the correct interpretation of events or the possibility of alternatives appear as less of an existential threat and more a source of bemusement or even as a means of constructing community among the interlocutors. Fontane's *Der Stechlin* provides some of the most striking examples. In addition to the modal adverbs, passages from this novel are replete with other linguistic markers of modality, especially subjunctive forms to mark hypothetical situations and verbs that indicate varying degrees of certainty ("meinen," "glauben," "wissen"). The novel, which relates, among other things, how the aging Major Dubslav von Stechlin runs for a position as member of the Reichstag and loses to a Social Democrat, is clearly concerned with the possibility of understanding a society in transition as well as of understanding those who represent the dynamic forces in that society. An example is the following exchange between two friends of Stechlin's son Woldemar, in which the epistemic elements appear in boldface:

> "Ja, das wollt ich. Ich wollte Sie nämlich wissen lassen, dass Ihr Célibataire seit Ausgang vorigen Winters in eben diesem Hause regelmäßig verkehrt."

> "Er wird **wohl** in vielen Häusern verkehren."

"**Möglich**, aber nicht sehr **wahrscheinlich**, da das eine Haus ihn ganz in Anspruch nimmt."

"Nun gut, so lassen wir ihn bei den Barbys. Aber **was bedeutet das**?"[35]

This rather quotidian exchange captures the general atmosphere of speculation that pervades these passages. Similar exchanges occur in the other novels as well, not infrequently in regard to the possibility of marriage across class lines, as in this somewhat longer example from *Frau Jenny Treibel*:

"Nein, Marcell, das letztere **gewiß** nicht; ich bin nicht für Zudringlichkeiten. Aber wenn Leopold morgen bei meinem Vater antritt - denn ich fürchte **beinah'**, daß er noch zu denen gehört, die sich, statt der Hauptperson, erst der Nebenpersonen versichern - wenn er also morgen antritt und um diese rechte Hand Deiner Cousine Corinna anhält, so nimmt ihn Corinna und fühlt sich als Corinne au Capitole."

"Das ist **nicht möglich**; Du **täuschest** Dich, Du spielst mit der Sache. Es ist eine Phantasterei, der Du nach Deiner Art nachhängst."

"Nein, Marcell, *Du* **täuschest** Dich, nicht ich; es ist mein vollkommener Ernst, so sehr, daß ich ein ganz klein wenig davor erschrecke."

"Das ist Dein Gewissen."

"**Vielleicht**. **Vielleicht** auch nicht. Aber so viel will ich Dir ohne weiteres zugeben, *das*, wozu der liebe Gott mich **so recht eigentlich** schuf, das hat nichts zu thun mit einem Treibel'schen Fabrikgeschäft, oder mit einem Holzhof und **vielleicht** am wenigsten mit einer Hamburger Schwägerin."[36]

---

[35] Right, I wanted to do that. I wanted to let you know that your celibate has, since the end of this past winter, been a regular guest in this very house.

[36] No Marcel, certainly not the latter; I don't go in for intrusions. But if Leopold should present himself tomorrow at my father's—for I almost fear he is one of those who still believes that he must start with the minor character rather than the protagonist—as I said if he should present himself tomorrow to ask for the right hand of your cousin Corinna, then Corinna will accept him and will think of herself as Corinna au Capitole.

It's impossible. You are deluding yourself, treating it like a game. It's one of those wild fantasies that you so often indulge in.

No, Marcel, *you* are deluding yourself, not me. I am being completely serious, so serious that I am even a bit frightened by the thought.

That is your conscience speaking.

Maybe. But then again maybe not. But this much I am happy to concede to you: the purpose for

The list of examples could be easily extended, and they remind us that we must beware of conflating structure and function. In all three of the authors discussed—Kafka, May, and Fontane—we find a high frequency of modal formulations such that the idea of epistemic narration seems quite reasonable. The extent to which these modal formulations serve the same function across these texts, however, remains an open question at this point. A case can be made that all of the novels represent attempts to manage epistemological uncertainty, and I think an even stronger case can be made for the more specific claim that psychological opacity, or the challenge of deciphering character, is the key (though not the only) driver of this uncertainty. But intensity of this uncertainty and the possibilities for its resolution differ substantially. In Kafka seemingly straightforward expressions of necessity tend to be undermined by other rhetorical features, whereas in May and Fontane expressions of mere possibility sometimes appear as false modesty on the part of characters who feel quite certain of the accuracy of their assertions. The tantalizing possibility raised by the above-average frequency of these modal expressions in all of these novels is that it might serve as the starting point for a delimiting of a broad and historically specific epistemic terrain. One of the scholarly commonplaces with regard to literary modernism is that it constitutes a response to an epistemological crisis, in the words of one scholar, "a crisis in 'knowing' precipitated in large part by nineteenth-century biology (Darwinism) and twentieth-century physics (quantum theory)."[37] To consider modernism not as an "abrupt collapse . . . of shared assumptions about the world" but as in intensification of epistemic positioning in all its forms, including doubt, possibility, *and* necessity or certainty, would require a significant reorientation in our thinking, both at the level of genre (Kafka *and* May) and of period (Kafka *and* Raabe).[38]

With possibility of such a reorientation, we circle back to the original question of historical context. Extending the arguments of Felski and Berman, we can claim that not only affect, but also epistemology serves as a conceptual category that enables us to consider the transtemporal resonance of literary works. To identify such categories, however, is not to argue against contextualization per se. As Felski herself insists, the fetishization of origins is not be countered with a naïve advocacy of the timelessness of art: "We cannot close our eyes to the historicity of artworks" (575). What is needed is a fresh approach to the multi-layered quality of that historicity, one that enables us to identify both the specificity of

---

which the good Lord actually put me on this earth, *that* has nothing to do with Treibel industries, or with a lumber yard, and perhaps least of all with a mother-in-law from Hamburg.

   [37] Jewel Spears Brooker, *Mastery and Escape: T. S. Eliot and the Dialectic of Modernism* (Amherst, MA: University of Massachusetts Press, 1994), 167.

   [38] Brooker, *Mastery and Escape.*

local (in a temporal sense) shifts and their connection to longer-term continuities. Corpus-based, computational approaches may help us to escape the box of synchronicity without lapsing into assertions of aesthetic autonomy, enabling us to begin to discern how the sources of the transtemporal appeal of the novel are always modulated by the historical moment.

Appendix: Seeded Topic Modeling

While the intuitive plausibility of the terms generated by this approach provided compelling support for the reliability of the method, we also ran two additional simple tests on the English corpus to confirm that we were not generating topics that were merely characteristic of non-fiction texts as opposed to novels, or of philosophical texts more generally conceived. In the first we re-ran the analysis using identical parameters but substituting a "random" non-fiction work for the epistemology document, Tocqueville's *Democracy in America*. In the second we swapped out the epistemology document with a similarly constructed ethics document and re-ran the analysis again. The results are captured in Table A.

Category	Document %	Top Words
Epistemology (Topic 32)	63.10%	mind reason things human truth knowledge true thing religion good natural life general make part virtue power opinion sense ideas
Democracy (Topic 24)	63.50%	people country power public state general law society english england part time war number order common present means influence
Ethics (Topic 23)	37.10%	case made mind time fact matter point present question make subject state general character interest good doubt circumstances part means
Ethics (Topic 42)	20.00%	god good religion life christian church things power spirit human true reason mind faith holy soul evil truth virtue heaven

Table A. Comparison of Results with Different Seed Texts

The lists of the top 20 words in Table A shed important light on the way in which the algorithms work and thereby on the strengths as well as the potential shortcomings of the approach. The algorithms generate "topics" in the form of a probability distribution over words, that is to say, a list of words that are likely to co-occur in a given document along with a ranking of how likely they are to co-occur. Keeping this definition in mind and reviewing the sets of terms, we can be reassured by the clear distinction between the epistemology terms and those that are most characteristic of Tocqueville's work. We do find limited overlap, but any passage that ranked highly for the *Democracy* topic 24 would presumably address society, politics, and public life to a much higher degree than would be the case for a passage that ranked highly for topic 32.

A comparison of the epistemology and the ethics topics presents us with a more complex situation. Most notably, the algorithms generated not just one but two significant semantic fields characteristic of the ethics document. In addition, both of these topics, and especially ethics topic 42, demonstrate substantial overlap with the epistemology topic. In fact, this complexity is exactly what we would expect. With regard to the split, we should remember that setting the number of topics is a way of influencing the extent to which individual topics map onto individual documents in the corpus. The fact that a 50 topic parameter leads to one epistemology topic but two ethics topic suggests that, when evaluated in the context of the corpus as a whole, the ethics document is less lexically distinct from the novels than the epistemology document - a plausible interpretation. Indeed, in light of the top words in each of the ethics topics, it would appear that one way to categorize the ethical preoccupations of the texts in the corpus is on the basis of their religious versus their juristic emphasis. With regard to the overlap, we should note that it pertains to terms that can be considered generally characteristic of philosophical texts—"human," "mind," "things," "reason," "true"—but we also find key terms unique to each the categories. The rank order of the terms is also crucial—the fact that "reason," "truth," knowledge," and "true" all appear in the top ten terms in the epistemology topic but not in either of the ethics topics.

It would be possible to use the topic modeling algorithms, or a less statistically sophisticated approach, to generate a restricted list of more "purely" epistemological terms, but such a list would be at odds with the way that epistemological concerns actually manifest in literary texts, which is certainly not in the form of a tightly-knit assemblage of the technical vocabulary typical of twentieth and twenty-first century epistemological treatises.[39] One expects novelistic epistemologies to be more diffuse and more intertwined with any number of other thematic engagements. Indeed, even in the case of philosophical texts, it can be exceedingly difficult to ascribe them to a single subdiscipline, especially in the case of works written in the seventeenth and eighteenth centuries. The "wider net" of terms generated by the topic modeling constitutes an advantage to the extent that it increases the likelihood of snaring epistemological passages in the novels that do not take the form of an explicit philosophical disquisition. The disadvantage is that it also creates the possibility of false positives, although the specificity of the *Democracy in America* topic should reassure us that these false

---

[39]By way of comparison: using *Voyant* to calculate the most frequently occurring terms in the *Stanford Encyclopedia of Philosophy* article on epistemology generates the following list of the top 25 (with word frequencies): justification (173); knowledge (147); know (131); beliefs (115); belief (106); biv (106); justified (102); according (83); perceptual (83); hands (70); experiences (60); coherentism (59); true (57); argument (54); evidence (53); epistemology (49); believing (48); question (48); basic (47); foundationalism (45); having (44); say (41); tim (40); reliability (38); way (37). The words "biv" and "tim" are abbreviations created by the author of the article to describe particular approaches.

positives will almost always be cases of philosophical passages whose epistemological content is less conspicuous, rather than passages on some totally unrelated area of concern.